Methods in Ecology and Evolution

British Ecological Society

*Methods in Ecology and Evolution* 2014, **5,** 1198–1205 doi: 10.1111/2041-210X.12261

APPLICATION

# ENMeval: An R package for conducting spatially independent evaluations and estimating optimal model complexity for MAXENT ecological niche models

**Robert Muscarella[1]*, Peter J. Galante[2], Mariano Soley-Guardia[2,3], Robert A. Boria[2], Jamie M. Kass[2,3], María Uriarte[1] and Robert P. Anderson[2,3,4]**

[1]*Department of Ecology, Evolution and Environmental Biology, Columbia University, 1200 Amsterdam Ave., New York, NY 10027, USA;* [2]*Department of Biology, City College of the City University of New York, 160 Convent Ave., New York, NY 10031, USA;* [3]*Graduate Center of the City University of New York, 365 5th Ave., New York, NY 10016, USA; and* [4]*Division of Vertebrate Zoology (Mammalogy), American Museum of Natural History, Central Park West & 79th Street, New York, NY 10024, USA*

## Summary

**1.** Recent studies have demonstrated a need for increased rigour in building and evaluating ecological niche models (ENMs) based on presence-only occurrence data. Two major goals are to balance goodness-of-fit with model complexity (e.g. by 'tuning' model settings) and to evaluate models with spatially independent data. These issues are especially critical for data sets suffering from sampling bias, and for studies that require transferring models across space or time (e.g. responses to climate change or spread of invasive species). Efficient implementation of procedures to accomplish these goals, however, requires automation.

**2.** We developed `ENMeval`, an R package that: (i) creates data sets for *k*-fold cross-validation using one of several methods for partitioning occurrence data (including options for spatially independent partitions), (ii) builds a series of candidate models using MAXENT with a variety of user-defined settings and (iii) provides multiple evaluation metrics to aid in selecting optimal model settings. The six methods for partitioning data are $n-1$ jackknife, random *k*-folds ( = bins), user-specified folds and three methods of masked geographically structured folds. `ENMeval` quantifies six evaluation metrics: the area under the curve of the receiver-operating characteristic plot for test localities ($AUC_{TEST}$), the difference between training and testing AUC ($AUC_{DIFF}$), two different threshold-based omission rates for test localities and the Akaike information criterion corrected for small sample sizes (AICc).

**3.** We demonstrate `ENMeval` by tuning model settings for eight tree species of the genus *Coccoloba* in Puerto Rico based on AICc. Evaluation metrics varied substantially across model settings, and models selected with AICc differed from default ones.

**4.** In summary, `ENMeval` facilitates the production of better ENMs and should promote future methodological research on many outstanding issues.

**Key-words:** ecological niche model, species distribution model, overfitting, model complexity, AIC, software, bioinformatics

## Introduction

Correlative ecological niche models (ENMs, often termed species distribution models, SDMs) have become an important tool for research in ecology, conservation and evolutionary biology (Guisan & Thuiller 2005; Elith *et al.* 2006; Kozak, Graham & Wiens 2008; Dormann *et al.* 2012). These tools, however, are subject to a number of methodological issues including the challenge of balancing goodness-of-fit with model complexity (Warren & Seifert 2011), and the need to evaluate model performance with independent data (Veloz 2009; Hijmans 2012).

A number of recent studies have demonstrated the sensitivity of ENM performance to model specification (e.g. Araújo & Guisan 2006; Elith, Kearney & Phillips 2010; Anderson & Gonzalez 2011; Elith *et al.* 2011; Warren & Seifert 2011; Araújo & Peterson 2012; Merow, Smith & Silander 2013; Shcheglovitova & Anderson 2013; Syfert, Smith & Coomes 2013; Radosavljevic & Anderson 2014; Warren *et al.* 2014). Two main conclusions are that: (i) species-specific tuning of settings (also called 'smoothing') can improve model performance, and (ii) spatially independent training and testing (also called

*Correspondence author. E-mail: bob.muscarella@gmail.com

'calibration' and 'evaluation') data sets can reduce the degree to which models are overfit (e.g. to biased sampling). These issues are particularly critical for studies involving transfer across space or time, especially those requiring extrapolation into non-analog conditions (e.g. Elith, Kearney & Phillips 2010; Anderson 2013). In practice, however, these recommendations are rarely implemented, largely because they can be prohibitively laborious and time-consuming (Phillips & Dudík 2008). As a result, most empirical studies rely on default settings of a given algorithm/software package and potentially biased evaluation methods.

We developed an R package (`ENMeval`) to help address these issues. Specifically, the current version of `ENMeval` facilitates construction and evaluation of ENMs with one of the most commonly used presence-only methods, Maxent (Phillips, Anderson & Schapire 2006). The structure of `ENMeval`, however, will allow later expansion to other niche modelling algorithms. Briefly, Maxent quantifies statistical relationships between predictor variables at locations where a species has been observed versus 'background' locations in the study region. These modelled relationships are constrained by various transformations of the original predictor variables ('feature classes' or FCs) – allowing more FCs enables more flexible and complex fits to the observed data. Higher flexibility, however, can increase the propensity for model overfitting (Peterson *et al.* 2011). By default, Maxent determines which FCs to allow based on the number of occurrence localities in a data set. Regardless of which feature classes are *permitted* in a model run, Maxent provides protection against overfitting via regularization, which penalizes the inclusion of additional parameters that result in little or no 'gain' to the model (Merow, Smith & Silander 2013). Users can specify which FCs to allow, and adjust the level of regularization via a single regularization multiplier (RM; default = 1·0). The RM acts in concert across all FCs as a coefficient multiplied to the individual regularization values (βs in Maxent) that correspond to each respective FC (Phillips & Dudík 2008). Several existing studies provide additional details on the mathematical underpinnings of Maxent (Phillips, Anderson & Schapire 2006; Phillips & Dudík 2008; Elith *et al.* 2011; Merow, Smith & Silander 2013; Yackulic *et al.* 2013).

Although the current default settings in Maxent were based on an extensive empirical tuning study (Phillips & Dudík 2008), recent work has shown that they can result in poorly performing models (Shcheglovitova & Anderson 2013; Radosavljevic & Anderson 2014). Additionally, artificial spatial autocorrelation between training and testing data partitions (e.g. due to sampling bias) can inflate metrics used to evaluate model performance (Veloz 2009; Wenger & Olden 2012; Radosavljevic & Anderson 2014). `ENMeval` should help address these issues and facilitate increased rigour in the development of Maxent models.

## Package description

`ENMeval` provides a number of novel resources for Maxent users. First, it includes six methods to partition data for training and testing, including three designed to achieve spatially independent splits. Secondly, it executes a series of models across a user-defined range of settings (i.e. combinations of FCs and RM values). Finally, it provides six evaluation metrics to characterize model performance. All of these operations can be completed with a single call to the primary function of the package, `ENMevaluate`, although supporting functions are also available (Table 1). The evaluation metrics returned can be used to compare models, and depending on the user's choice of evaluation criteria, select the optimally performing settings. `ENMeval` specifically does not perform model selection because it is not clear which optimality criteria are most appropriate for evaluating ENMs (Fielding & Bell 1997; Lobo, Jiménez-Valverde & Real 2008; Peterson *et al.* 2011; Warren & Seifert 2011). Rather, the various evaluation statistics provided can be used to select settings based on recommendations from current and future literature. Below, we briefly outline the components of the package and demonstrate its functionality by conducting species-specific tuning for eight species of native trees in Puerto Rico.

### DATA PARTITIONING AND MODEL EXECUTION

A run of `ENMevaluate` begins by using one of the six methods to partition occurrence localities into testing and

**Table 1.** The functions included in `ENMeval`. See the main text and package manual (Appendix S2) for additional details

| Function name | Description |
|---|---|
| calc.aicc | Calculate the Akaike Information Criterion corrected for small samples sizes (AICc) based on Warren & Seifert (2011) |
| calc.niche.overlap | Compute pairwise niche overlap (similarity of estimated suitability scores) in geographic space for Maxent predictions. The value ranges from 0 (no overlap) to 1 (identical predictions). Based on the 'nicheOverlap' function of the dismo R package (Hijmans *et al.* 2011) |
| corrected.var | Calculates variance corrected for non-independence of *k*-fold iterations (Shcheglovitova & Anderson 2013) |
| `ENMevaluate` | The primary function of `ENMeval`, this function automatically executes Maxent across a range of feature class and regularization multiplier settings, providing several evaluation metrics to aid in identifying settings that balance model goodness-of-fit with model complexity |
| get.evaluation.bins | A general title for six separate functions ('get.randomkfold', 'get.jackknife', 'get.user', 'get.block', 'get.checkerboard1', 'get.checkerboard2') that partition occurrence and background localities into separate bins for training and testing (i.e. calibration and evaluation) |
| make.args | Generate a list of arguments to pass to Maxent and to use as labels in plotting |
| eval.plot | A basic plotting function to visualize evaluation metrics generated by `ENMevaluate` |

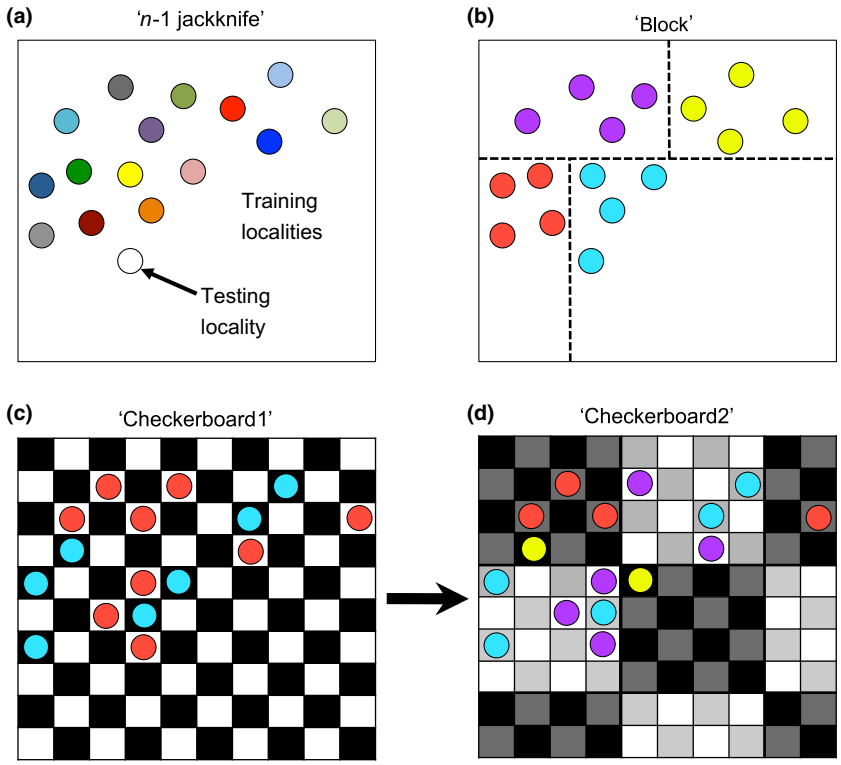


**Fig. 1.** `ENMeval` provides six methods for partitioning data into bins (each of which implements a variation on *k*-fold cross-validation), four of which are illustrated here. In all panels, different coloured occurrence localities represent different bins. (a) With the *n*−1 jackknife method, each of *n* occurrence localities is used for testing once (e.g. white locality here), while all others are used for training in that iteration (coloured localities). A total of *n* models are run, and evaluation metrics are summarized across these iterations. (b) The 'block' method partitions data into four bins based on the lines of latitude and longitude that divide occurrence localities as equally as possible. The amount of geographic (and environmental) space corresponding to each bin, however, is likely to differ. (c and d) The two 'checkerboard' methods involve aggregating the original environmental input grids into either one or two checkerboard-like grids based on user-defined aggregation factors. For the 'checkerboard1' method (c), a single grid is used to partition occurrence localities into two bins. The 'checkerboard2' method (d) is identical to 'checkerboard1' except that an additional second level of spatial aggregation is specified (i.e. fine- and coarse-grain aggregation). Localities are first partitioned into two groups according to the 'fine'-grain checkerboard (as in 'checkerboard1'). Then, these groups are further subdivided into two groups each based on the 'coarse'-grain checkerboard (with an aggregation factor specified by the user), yielding 4 bins. Note that, creating these grids to define bins does not affect the grain size of the environmental predictor variables themselves. As opposed to the 'block' method, both 'checkerboard' methods result in an approximately equal sample of geographic (and likely environmental) space in each bin. The numbers of occurrence localities, however, are likely to differ across bins. A warning is given if fewer than four bins are created as a result of the spatial configuration of occurrence localities. For each of these last three methods (b, c and d), *k* models are run iteratively using *k*−1 bins for training and the remaining one for testing. Evaluation metrics are then summarized across the *k* iterations.

training bins (folds) for *k*-fold cross-validation (Fig. 1; Fielding & Bell 1997; Peterson *et al.* 2011). The 'random *k*-fold' method partitions occurrence localities randomly into a user-specified number of (*k*) bins (equivalent to the 'cross-validate' partitioning scheme available in the current version of the MAXENT software). Primarily when working with small data sets (e.g. < ca. 25 localities), users may choose a special case of *k*-fold cross-validation where the number of bins (*k*) is equal to the number of occurrence localities (*n*) in the data set (Fig. 1a; Pearson *et al.* 2007; Shcheglovitova & Anderson 2013). Note that neither of these methods accounts for spatial autocorrelation between testing and training localities, which can inflate evaluation metrics, at least for data sets that result from biased sampling (Veloz 2009; Hijmans 2012; Wenger & Olden 2012). As a third option, users can define *a priori* partitions, which provides a flexible way to conduct spatially independent cross-validation with background masking (see below).

Three additional methods are variations of what Radosavljevic & Anderson (2014) referred to as 'masked geographically structured' data partitioning (Fig. 1). The 'block' method partitions data according to the latitude and longitude lines that divide the occurrence localities into four bins of (insofar as possible) equal numbers (Fig. 1b). Both occurrence and background localities are assigned to each of the four bins based on their position with respect to these lines. `ENMeval` also includes two variants of a 'checkerboard' approach to partition occurrence localities. These generate checkerboard grids across the study extent that partition the localities into bins (Fig. 1c,d). In contrast to the block method, the checkerboard methods subdivide geographic space equally but do not ensure a balanced number of occurrence localities in each bin.

Choosing among the data partitioning methods depends on the research objectives and the characteristics of the study system. For example, the block method may be desirable for

**Table 2.** The evaluation metrics calculated by `ENMeval`

| Metric | Description | References |
|---|---|---|
| $AUC_{TEST}$ | The threshold-independent metric AUC based on predicted values for the test localities (i.e. localities withheld during model training), averaged over *k* iterations. Higher values reflect a better ability for a model to discriminate between conditions at withheld (testing) occurrence localities and those of background localities (by ranking the former higher than the latter based on their predicted suitability values). The rank-based AUC does not indicate model fit | Hanley & McNeil (1982), Peterson *et al.* (2011) |
| $AUC_{DIFF}$ | The difference between the AUC value based on training localities (i.e. $AUC_{TRAIN}$) and $AUC_{TEST}$ ($AUC_{TRAIN} - AUC_{TEST}$). If $AUC_{TRAIN} < AUC_{TEST}$, the returned value is zero. Value of $AUC_{DIFF}$ is expected to be positively associated with the degree of model overfitting | Warren & Seifert (2011) |
| $OR_{MTP}$ ('Minimum Training Presence' omission rate) | A threshold-dependent metric that indicates the proportion of test localities with suitability values (MAXENT relative occurrence rates) lower than that associated with the lowest-ranking training locality. Omission rates greater than the expectation of zero typically indicate model overfitting | Fielding & Bell (1997), Peterson *et al.* (2011), Radosavljevic & Anderson (2014) |
| $OR_{10}$ (10% training omission rate) | A threshold-dependent metric that indicates the proportion of test localities with suitability values (MAXENT relative occurrence rates) lower than that excluding the 10% of training localities with the lowest predicted suitability. Omission rates greater than the expectation of 10% typically indicate model overfitting | Fielding & Bell (1997), Peterson *et al.* (2011) |
| AICc | The Akaike Information Criterion corrected for small samples sizes reflects both model goodness-of-fit and complexity. The model with the lowest AICc value (i.e. $\Delta AICc = 0$) is considered the best model out of the current suite of models; all models with $\Delta AICc < 2$ are generally considered to have substantial support | Burnham & Anderson (2004), Warren & Seifert (2011) |

studies involving model transfer across space or time, including the possibility of encountering non-analog conditions (e.g. native versus invaded regions, climate change effects; Wenger & Olden 2012). In contrast, the checkerboard methods (which are less likely to require extrapolation in environmental space) may be more appropriate when model transferability is not required. Nonetheless, we emphasize that evaluating models with various combinations of data partitions and software settings does not guarantee the reliability of models projected across space or time. For applications that rely on model transferability, researchers should identify non-analog conditions, illustrate extrapolated response curves, quantify uncertainty based on the manner of extrapolation and interpret those predictions with additional caution.

After data partitioning, the `ENMevaluate` function iteratively builds *k* models for each combination of settings, *s*,

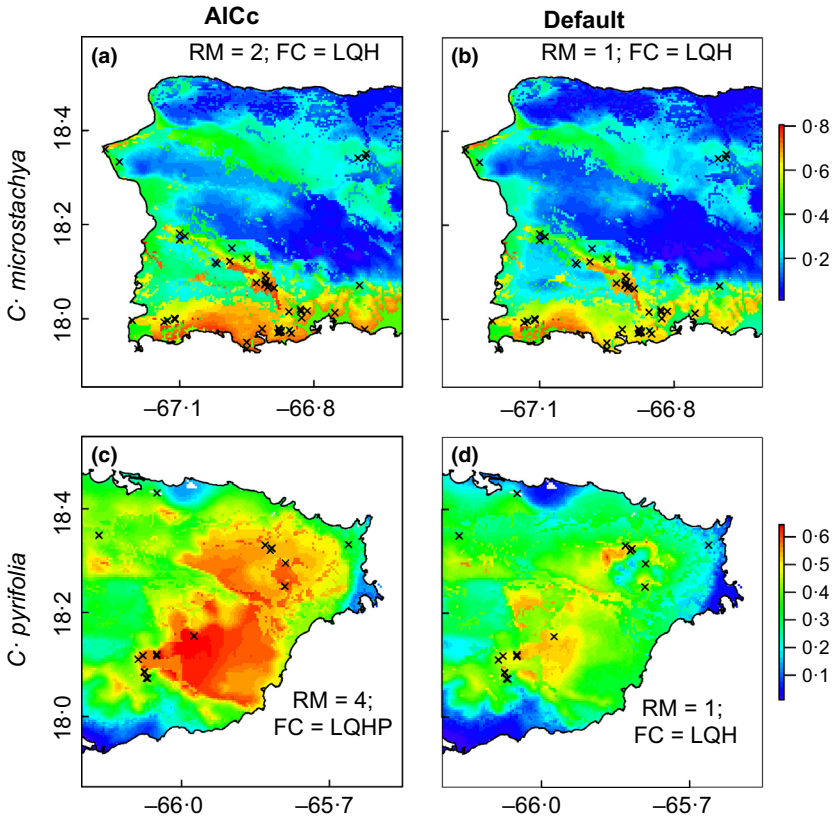


**Fig. 2.** Occurrences and predictions of MAXENT models shown in selected portions of Puerto Rico for *Coccoloba microstachya* (a, b) and *C. pyrifolia* (c, d). Models shown here correspond to those producing the minimum AICc (a, c), and those built with default settings (b, d). Occurrence localities are indicated with an x. Scale bars show MAXENT logistic output (used here only for visualization purposes); higher values (warmer colours) indicate higher predicted suitability.

using $k-1$ bins for model training and the withheld bin for testing. Importantly, for the geographically structured partitioning methods, background localities in the same geographic area as the bin holding testing localities are not included in the training phase (Phillips 2008; Phillips & Dudík 2008; Radosavljevic & Anderson 2014). A 'full' model (using the entire, unpartitioned data set) is also made to calculate AICc, resulting in a total of $s \times (k + 1)$ model runs.

### EVALUATION METRICS

Because no consensus currently exists regarding the most appropriate metric or approach to evaluate performance of ENMs (Fielding & Bell 1997; Lobo, Jiménez-Valverde & Real 2008; Peterson *et al.* 2011; Warren & Seifert 2011), `ENMeval` provides several metrics likely to be useful for presence-background evaluations (Table 2). All calculations in `ENMeval` are based on MAXENT raw output values (Merow, Smith & Silander 2013; Yackulic *et al.* 2013). Note, however, that any rescaling that preserves rank (e.g. cumulative or logistic) will lead to the same evaluation values for the rank-based metrics used here (based on omission rates or AUC), but not for AICc (Warren *et al.* 2009; Peterson *et al.* 2011) or Schoener's *D* (see below). First, `ENMeval` calculates a measure of the model's ability to discriminate conditions at withheld occurrence localities from those at background samples: the area under the curve of the receiver operating characteristic plot based on the testing data ($AUC_{TEST}$). In this implementation, $AUC_{TEST}$ is calculated with the full set of background localities (corresponding to all $k$ bins) to enable comparison among $k$-fold iterations (Radosavljevic & Anderson 2014). Secondly, to quantify overfitting, `ENMeval` calculates the difference between training and testing AUC ($AUC_{DIFF}$), which is expected to be high with overfit models (Warren & Seifert

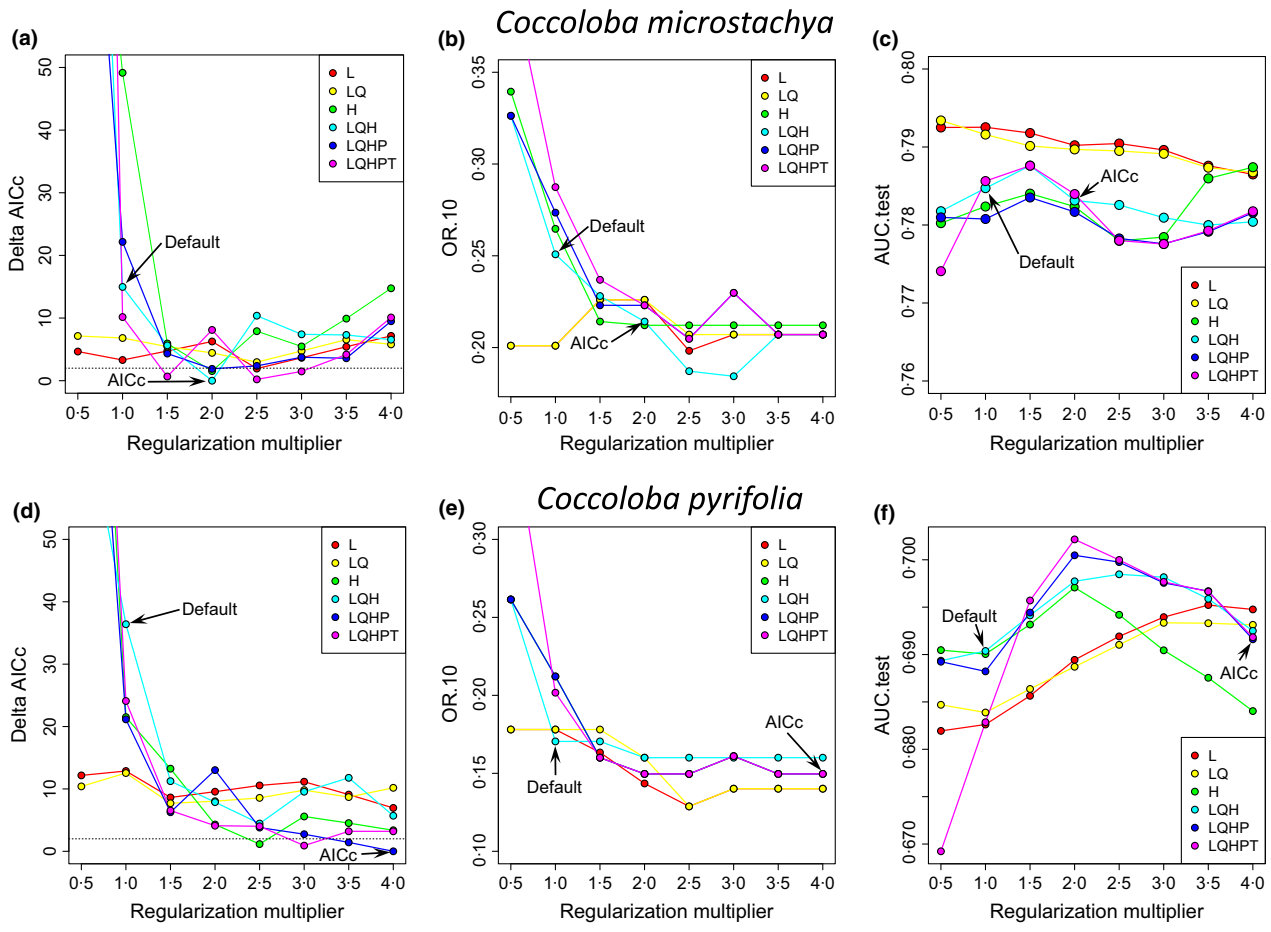


**Fig. 3.** Three evaluation metrics for *Coccoloba microstachya* (a–c) and *C. pyrifolia* (d–f) resulting from MAXENT models made across a range of feature-class combinations and regularization multipliers. Left panels show ΔAICc, centre panels show the omission rate of testing localities at the 10% training threshold ($OR_{10}$), and right panels show $AUC_{TEST}$. Dotted horizontal line in ΔAICc plots represents ΔAICc = 2, which delimits models that are generally considered to have substantial support relative to others examined – that is those below the line (Burnham & Anderson 2004). Default settings and settings that yielded minimum AICc are indicated with arrows. Legends denote feature classes allowed (L = linear, Q = quadratic, H = hinge, P = product and T = threshold). Note that for these species, AICc-selected settings (based on all localities) resulted in substantially lower omission rates (in the models run with the partitioned data; 'checkerboard2' method) than were achieved by the default settings. However, maximal $AUC_{TEST}$ showed low correspondence with either the AICc-chosen or default settings. For these species, AICc consistently selected regularization multipliers higher than the default value. Results for all eight species (including estimates of variance) are provided in Appendix S1.

**Table 3.** Evaluation metrics of MAXENT ENMs generated by `ENMeval` for eight species of *Coccoloba* trees in Puerto Rico

| Species | *n* | Model | FC | RM | $AUC_{TEST}$ | $AUC_{DIFF}$ | $OR_{MTP}$ | $OR_{10}$ | ΔAICc | *D* |
|---|---|---|---|---|---|---|---|---|---|---|
| *C. costata* (C. Wright) | 48 | AICc | LQHPT | 1·5 | 0·823 | 0·070 | 0·039 | 0·175 | 0·0 | 0·932 |
| | | Default | LQH | 1 | 0·829 | 0·069 | 0·062 | 0·175 | 20·9 | |
| *C. diversifolia* (Jacq.) | 69 | AICc | LQ | 3 | 0·760 | 0·080 | 0·081 | 0·222 | 0·0 | 0·807 |
| | | Default | LQH | 1 | 0·776 | 0·078 | 0·056 | 0·250 | 95·8 | |
| *C. krugii* (Lindau) | 24 | AICc | L | 3 | 0·958 | 0·017 | 0·071 | 0·117 | 0·0 | 0·849 |
| | | Default | LQH | 1 | 0·945 | 0·030 | 0·071 | 0·242 | 19·9 | |
| *C. microstachya* (Willd.) | 58 | AICc | LQH | 2 | 0·783 | 0·070 | 0·098 | 0·214 | 0·0 | 0·916 |
| | | Default | LQH | 1 | 0·785 | 0·073 | 0·023 | 0·251 | 15·0 | |
| *C. pyrifolia* (Desf.) | 71 | AICc | LQHP | 4 | 0·692 | 0·053 | 0·022 | 0·150 | 0·0 | 0·871 |
| | | Default | LQH | 1 | 0·690 | 0·083 | 0·010 | 0·170 | 36·4 | |
| *C. sintenisii* (Urb.) | 27 | AICc | LQ | 1·5 | 0·798 | 0·090 | 0·069 | 0·188 | 0·0 | 0·822 |
| | | Default | LQH | 1 | 0·813 | 0·097 | 0·091 | 0·272 | NA | |
| *C. swartzii* (Meisn.) | 42 | AICc | L | 1·5 | 0·713 | 0·092 | 0·019 | 0·128 | 0·0 | 0·909 |
| | | Default | LQH | 1 | 0·697 | 0·133 | 0·128 | 0·300 | 29·8 | |
| *C. venosa* (L.) | 44 | AICc | LQH | 2·5 | 0·712 | 0·113 | 0·095 | 0·312 | 0·0 | 0·860 |
| | | Default | LQH | 1 | 0·707 | 0·145 | 0·115 | 0·312 | 53·9 | |

Results are based on the 'checkerboard2' method for data partitioning and are shown for settings that gave minimum AICc values (i.e. ΔAICc = 0) as well as for MAXENT default settings. Number of occurrence records used for each species is given by *n*. Schoener's *D* statistic (Schoener 1968; Warren *et al.* 2009) – a measure of model similarity in geographic space – compares the predictions of AICc-selected models with those based on default settings. Values of *D* range from zero to one, with higher values indicating more similar models.

2011). `ENMeval` also calculates two omission rates that quantify model overfitting when compared with the respective theoretically expected omission rates: the proportion of test localities with MAXENT output values lower than that corresponding to (i) the training locality with the lowest value (i.e. the minimum training presence, MTP; = 0% training omission) or (ii) a 10% omission rate of the training localities ( = 10% training omission) (Pearson *et al.* 2007). `ENMeval` provides the mean and variance (corrected for the non-independence of the *k* iterations, Shcheglovitova & Anderson 2013) for each of these four metrics. The function also calculates the AICc value, ΔAICc and AICc weight for each full model, providing information on the relative quality of a model given the data (Burnham & Anderson 2004; Warren & Seifert 2011). Because AICc is calculated using the full data set, it is not affected by the method chosen for data partitioning. We note that, following Warren & Seifert (2011), AICc is calculated based on the number of non-zero parameters in the MAXENT lambda file and that this value may not accurately estimate the total degrees of freedom in the model (Hastie, Tibshirani & Friedman 2009). Nonetheless, the relative performance of AICc versus other model evaluation metrics is a topic of ongoing research (e.g. Cao *et al.* 2013; Radosavljevic & Anderson 2014; Warren *et al.* 2014), and `ENMeval` should help advance this line of research. Finally, to quantify how predictions differ in geographic space (e.g. Fig. 2), `ENMeval` computes pairwise niche overlap between models using Schoener's *D* (Schoener 1968; Warren *et al.* 2009). Finally, `ENMeval` includes a basic plotting function to visualize evaluation statistics (see Fig. 3).

Recent work has demonstrated equivalency between the MAXENT algorithm and loglinear generalized linear models (Renner & Warton 2013), as well as close links to inhomogeneous Poisson process (IPP) models (Fithian & Hastie 2013). These connections open numerous additional diagnostic tools that are not readily available in the current MAXENT software (e.g. using the data to determine the most appropriate spatial resolution of predictor variables). Future work will benefit by capitalizing on the connections among these approaches.

## Case study

To demonstrate `ENMeval`, we tuned MAXENT models for eight native tree species from the genus *Coccoloba* (Polygonaceae) in Puerto Rico (Table 3). We compiled occurrence localities (ranging from 24 to 71 across species) from herbaria at the University of Puerto Rico, the US National Museum of Natural History and the New York Botanical Garden. As predictor variables, we used a categorical map of soil parent material (Bawiec 1999) and four climatic variables: log-transformed mean annual precipitation (log [mm $year^{-1}$]), coefficient of variation of monthly mean precipitation, mean temperature of the coldest month (°C) and mean daily temperature range (°C) (Daly, Helmer & Quiñones 2003). To reduce the influence of spatial sampling bias, we applied a weighted-target group approach (Anderson 2003) by using 22,858 tree species occurrence localities throughout Puerto Rico as background localities (Phillips *et al.* 2009). After partitioning occurrence data using the checkerboard2 method (see Fig. 1, aggregation factor = 5), we built models with RM values ranging from 0·5 to 4·0 (increments of 0·5) and with six different FC combinations (L, LQ, H, LQH, LQHP, LQHPT; where L = linear, Q = quadratic, H = hinge, P = product and T = threshold); this resulted in 1920 individual model runs.

Here, we summarize the relative performance of models built with default settings versus those selected by AICc (i.e. ΔAICc = 0; Fig. 2); comprehensive results are provided in Appendix S1. Settings of AICc-selected models differed from default settings for all species. Although we did not find general trends regarding FCs, AICc-selected models had higher RM

 

values than the default of 1·0 (Table 3; Fig. 3). Because higher regularization 'smoothes' response curves by imposing a higher penalty for the inclusion of parameters, this result suggests that default settings tended to result in more complex models relative to AICc-selected models (Fig. 2). Overall, AICc-selected models also had lower omission rates than default models, indicating less overfitting (Fig. 3). However, AICc-selected models generally showed slightly lower $AUC_{TEST}$ values than those made by default, suggesting somewhat lower discriminatory ability (Table 3; Fig. 3). This preliminary analysis reinforces the importance of species-specific tuning to build ENMs with Maxent.

## Conclusions

By relieving some of the logistic challenges associated with species-specific tuning and model evaluation, `ENMeval` facilitates research in ecological niche modelling. For instance, although beyond the scope of this software description, future work based on both simulated data sets and a variety of real species should help clarify the strengths and weaknesses of various data partitioning methods and evaluation metrics. Overall, we anticipate that `ENMeval` will help to advance research in model evaluation and methods for extrapolating ENMs in environmental space. Similar issues exist for algorithms other than Maxent, and the current structure of `ENMeval` will allow later incorporation of other algorithms.

## Obtaining `ENMeval`

`ENMeval` requires a current R installation (freely available from http://cran.r-project.org/) and can be downloaded from CRAN at: http://cran.r-project.org/web/packages/ENMeval/index.html. The package manual is provided in Appendix S2. `ENMeval` is under development and we welcome bug reports and feedback, including suggestions for features that could be included in future versions.

## Acknowledgements

We thank Fabiola Areces, Franklin Axelrod, Danilo Chinea and Jeanine Vélez at the University of Puerto Rico for providing access to digitized herbarium records. Rebecca Panko and Víctor José Vega López contributed by tirelessly georeferencing herbarium specimens. This manuscript benefitted from the comments of two anonymous reviewers. This research was supported by NSF-DEB 1311367 to MU and RM, NSF-DEB 1119915 to RPA, the Graduate Center of the City University of New York (Science Fellowship and Doctoral Student's Council Dissertation Award to MSG, CUNY Science Scholarship and Graduate Assistantship to JMK), and the Luis Stokes Alliance for Minority Participation (Bridge to Doctorate Fellowship to RAB). The authors have no conflict of interest to declare.

## Data accessibility

We compiled the occurrence data used in our case study from publicly accessible data bases including GBIF (www.gbif.org) and digitized records from several herbaria (NY, MAPR, US and UPRRP). Environmental data layers were compiled from published studies. Specifically, the geologic substrate layer comes from Bawiec (1999) and the climatic data layers come from Daly, Helmer & Quiñones (2003). We provide all necessary data in Appendix S3, together with an R script to repeat our case study analysis.

## References

Anderson, R.P. (2003) Real vs. artefactual absences in species distributions: tests for *Oryzomys albigularis* (Rodentia: Muridae) in Venezuela. *Journal of Biogeography*, **30**, 591–605.

Anderson, R.P. (2013) A framework for using niche models to estimate impacts of climate change on species distributions. *Annals of the New York Academy of Sciences*, **1297**, 8–28.

Anderson, R.P. & Gonzalez, J.I. (2011) Species-specific tuning increases robustness to sampling bias in models of species distributions: an implementation with Maxent. *Ecological Modelling*, **222**, 2796–2811.

Araújo, M.B. & Guisan, A. (2006) Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, **33**, 1677–1688.

Araújo, M. & Peterson, A.T. (2012) Uses and misuses of bioclimatic envelope modelling. *Ecology*, **93**, 1527–1539.

Bawiec, W.J., ed. (1999) Geology, geochemistry, geophysics, mineral occurrences and mineral resource assessment for the Commonwealth of Puerto Rico: U.S. Geological Survey Open-File Report 98-038, available online only.

Burnham, K.P. & Anderson, D.R. (2004) Multimodel inference: understanding AIC and BIC in Model Selection. *Sociological Methods and Research*, **33**, 261–304.

Cao, Y., DeWalt, R.E., Robinson, J.L., Tweddale, T., Hinz, L. & Pessino, M. (2013) Using Maxent to model the historic distributions of stonefly species in Illinois streams: the effects of regularization and threshold selections. *Ecological Modelling*, **259**, 30–39.

Daly, C., Helmer, E.H. & Quiñones, M. (2003) Mapping the climate of Puerto Rico, Vieques and Culebra. *International Journal of Climatology*, **23**, 1359–1381.

Dormann, C.F., Schymanski, S.J., Cabral, J., Chuine, I., Graham, C., Hartig, F. *et al.* (2012) Correlation and process in species distribution models: bridging a dichotomy. *Journal of Biogeography*, **39**, 2119–2131.

Elith, J., Kearney, M. & Phillips, S. (2010) The art of modelling range-shifting species. *Methods in Ecology and Evolution*, **1**, 330–342.

Elith, J., Graham, C.H., Anderson, R.P., Dudík, M., Ferrier, S., Guisan, A. *et al.* (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129–151.

Elith, J., Phillips, S.J., Hastie, T., Dudík, M., Chee, Y.E. & Yates, C.J. (2011) A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, **17**, 43–57.

Fielding, A.H. & Bell, J.F. (1997) A review of methods for the assessment of prediction errors in conservation presence-absence models. *Environmental Conservation*, **24**, 38–49.

Fithian, W. & Hastie, T. (2013) Finite-sample equivalence in statistical models for presence-only data. *Annals of Applied Statistics*, **7**, 1917–1939.

Guisan, A. & Thuiller, W. (2005) Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, **8**, 993–1009.

Hanley, J.A. & McNeil, B.J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**, 29–36.

Hastie, T., Tibshirani, R. & Friedman, J.H. (2009) *The elements of statistical learning: data mining, inference, and prediction*, 2nd edn. Springer-Verlag, New York, New York, USA.

Hijmans, R.J. (2012) Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model. *Ecology*, **93**, 679–688.

Hijmans, R.J., Phillips, S., Leathwick, J. & Elith, J. (2011) Package 'dismo'. Available online at: http://cran.r-project.org/web/packages/dismo/index.html.

Kozak, K.H., Graham, C.H. & Wiens, J.J. (2008) Integrating GIS-based environmental data into evolutionary biology. *Trends in Ecology and Evolution*, **23**, 141–148.

Lobo, J.M., Jiménez-Valverde, A. & Real, R. (2008) AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology & Biogeography*, **17**, 145–151.

Merow, C., Smith, M. & Silander, J.A. (2013) A practical guide to Maxent: what it does, and why inputs and settings matter. *Ecography*, **36**, 1–12.

Pearson, R.G., Raxworthy, C.J., Nakamura, M. & Peterson, A.T. (2007) Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. *Journal of Biogeography*, **34**, 102–117.

Peterson, A.T., Soberón, J., Pearson, R.G., Anderson, R.P., Martínez-Meyer, E., Nakamura, M. *et al.* (2011) *Ecological Niches and Geographic Distributions.* Monographs in Population Biology, 49. Princeton University Press, Princeton, New Jersey, USA.

Phillips, S. (2008) Transferability, sample selection bias and background data in presence-only modelling: a response to Peterson et al. (2007). *Ecography*, **31**, 272–278.

Phillips, S.J., Anderson, R.P. & Schapire, R.E. (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259.

Phillips, S.J. & Dudík, M. (2008) Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography*, **31**, 161–175.

Phillips, S.J., Dudík, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J. *et al*. (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications*, **19**, 181–197.

Radosavljevic, A. & Anderson, R.P. (2014) Making better Maxent models of species distributions: complexity, overfitting and evaluation. *Journal of Biogeography*, **41**, 629–643.

Renner, I.W. & Warton, D.I. (2013) Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology. *Biometrics*, **69**, 274–281.

Schoener, T.W. (1968) The *Anolis* lizards of Bimini: resource partitioning in a complex fauna. *Ecology*, **49**, 704–726.

Shcheglovitova, M. & Anderson, R.P. (2013) Estimating optimal complexity for ecological niche models: a jackknife approach for species with small sample sizes. *Ecological Modelling*, **269**, 9–17.

Syfert, M.M., Smith, M.J. & Coomes, D.A. (2013) The effects of sampling bias and model complexity on the predictive performance of MaxEnt species distribution models. *PLoS ONE*, **8**, e55158.

Veloz, S.D. (2009) Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. *Journal of Biogeography*, **36**, 2290–2299.

Warren, D.L. & Seifert, S.N. (2011) Ecological niche modeling in Maxent: the importance of model complexity and the performance of model selection criteria. *Ecological Applications*, **21**, 335–342.

Warren, D.L., Glor, R.E., Turelli, M. & Funk, D. (2009) Environmental niche equivalency versus conservatism: quantitative approaches to niche evolution. *Evolution*, **62**, 2868–2883; Erratum: Evolution 2865: 1215.

Warren, D.L., Wright, A.N., Seifert, S.N. & Shaffer, H.B. (2014) Incorporating model complexity and spatial sampling bias into ecological niche models of climate change risks faced by 90 California vertebrate species of concern. *Diversity and Distributions*, **20**, 334–343.

Wenger, S.J. & Olden, J.D. (2012) Assessing transferability of ecological models: an underappreciated aspect of statistical validation. *Methods in Ecology and Evolution*, **3**, 260–267.

Yackulic, C.B., Chandler, R., Zipkin, E.F., Royle, J.A., Nichols, J.D., Campbell Grant, E.H. *et al*. (2013) Presence-only modelling using MAXENT: when can we trust the inferences? *Methods in Ecology and Evolution*, **4**, 236–243.

*Received 16 June 2014; accepted 28 August 2014*
*Handling Editor: Jana McPherson*

## Supporting Information

Additional Supporting Information may be found in the online version of this article.

**Appendix S1.** Results for model tuning experiment for eight native tree species from Puerto Rico.

**Appendix S2.** Package manual for '`ENMeval`'.

**Appendix S3.** Data and R script used for case study.